単語埋め込みベクトルの圧縮法

単語間の意味的な関係をコンピュータ上で上手に扱う方法論として「単語埋め込みベクトル」と呼ばれる技術が注目されています．単語埋め込みベクトルを用いることで，例えば，単語間の意味的な関係を類推する演算に関して，より人間の感覚に近い結果が得られるようになりました．本稿では，単語埋め込みベクトル利用時の利便性を高めるために，意味関係に関する演算の精度を保ったまま必要記憶容量を大幅に削減する方法を紹介します．

すずき　じゅん　ながた　まさあき
鈴木　潤　/永田　昌明

NTTコミュニケーション科学基礎研究所

単語の意味を「演算」する

人間が扱う自然言語を，コンピュータが人間と同じように取り扱えるようになるには，克服すべき困難な課題がまだ多く残されています．特に，文章や単語間の意味的な関係を正しく解釈するのは非常に難しい課題です．そのような状況の中で，単語間の意味的な関係をコンピュータによる演算により扱う方法論として，「単語埋め込みベクトル」と呼ばれる技術が注目を集めています（**図1**）．この技術自体は，1980年代から存在する古典的な技術[(1)]ですが，ニューラルネットワークと同じように，近年の深層学習技術の急速な発展の中で改めて注目された技術といえます．具体的には，2013年に発表された論文[(2)]で，大規模データに対して非常に高速かつ精度良く構築できることが示されたことをきっかけに世界的に注目を集め，多くの発展研究がその後次々発表されました．また，現在も継続的に研究が進められている技術です．

1つの分かりやすい例として，「単語埋め込みベクトル」を利用すること

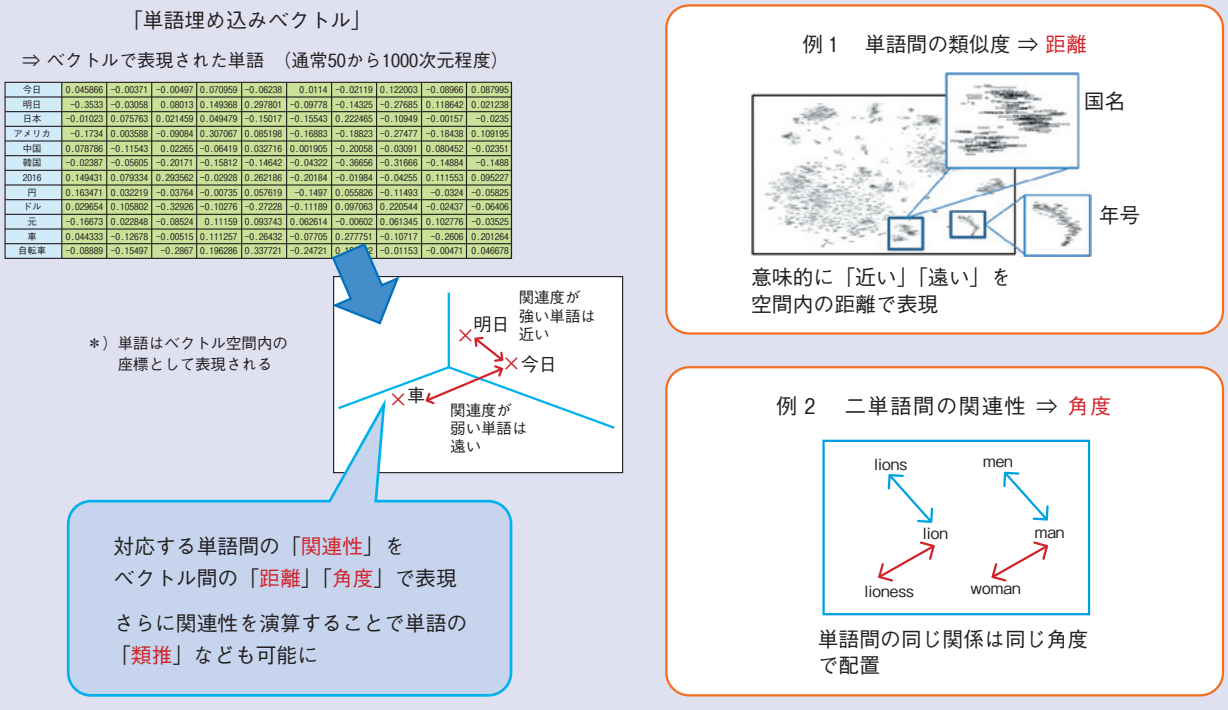


図1　単語埋め込みベクトル

で，単語を使った意味的な類推をベクトルの演算で実現することができます．例えば，「“フランス”に対する“ワイン”と同じ関係になると思う，“ドイツ”に対する語は？」という質問をした場合，多くの人は「ビール」と答えてくれます．もちろんこの質問に対する「唯一の正解」はありませんし「ビール」が正解とは限りません．しかし，多くの人が「ビール」という答えは妥当な回答だと感じると思います．この例から分かる重要な点は，こういった人間の直感や知識に近い答えを，コンピュータが演算で求めることができる時代になっている，ということです．

従来，こういった意味的な演算は，人手により構築されたコンピュータ用の辞書などを用いることで，ある程度は実現することができていました．しかし，単語埋め込みベクトルが従来の人手による辞書とは決定的に違う点は，網羅性にあります．容易に想像がつくこととして，辞書の場合は，人手により記述があればそれを使って比較的高度な意味処理が可能となりますが，記載がなければ全く対応することができません．また，人手による対応になるため，更新が難しく扱える語彙数や意味的関係は非常に限定的になります．例えば，前述の「ビール」の例は，辞書に記載される事柄としての優先度はかなり低いと考えられるため，従来の辞書を用いる方法では，解決できなかったと考えられます．一方，単語埋め込みベクトルは，大量の文章データから機械学習法を用いて自動的に構築します．原理的には，学習時に利用した文章に出現したすべての語彙を扱えます．実際に，現時点で数百万語彙でも何の問題もなく計算することができています．つまり，人手による辞書よりも圧倒的に多い語彙数で，かつ，それらの語彙間のすべての関係を網羅することができます（**図2**）．

コンピュータ・開発者にとって使いやすい単語埋め込みベクトル

このような技術を活用することで，対話，質問応答，情報検索，翻訳，要約といった，自然言語を介するあらゆる情報処理システムの中で，人間の感覚に近い意味的な処理を実現できるようになり，より自然で違和感のないシステムを実現することができると考えられます．このことから，今後のAI（人工知能）技術の発展を支える最重要技術の1つともとらえることができます．ただし，実際に単語埋め込みベクトルを実システム内で利用することを考えると，いくつかの不便な点がみられます．例えば，単語埋め込みベクトルの学習にはさまざまな乱数要素が存在するため，得られる単語埋め込みベクトルが学習するたびに毎回異なり再現性が低いという問題があります．また，単語埋め込みベクトルの次元数は事前に決定して構築しなくてはいけないため，さまざまな利用場面でシステムに必要な次元数が違うために毎回つくり直す必要があり，再利用性が低いという問題などが挙げられます．性能の良い最先端技術が実システムで簡単に利用できるとは限らないというのはよくあることだと思います．

そこで，私たちはこの最重要技術に対して，ここで挙げた学習の再現性や再利用性が低いといった使いにくい観点を解消していくことで，システム開発者が実際にシステムを構築する際の利便性を高めた単語埋め込みベクトルを構築する一連の技術を考案してきました[(3)～(5)]．本稿では，その中の1つとして，単語埋め込みベクトルを利用する際に必要となる記憶容量（メモリ量）を大幅に削減できる技術を紹介します[(5)]．

必要記憶容量の削減法

■必要記憶容量を削減することの意義

まず必要記憶容量を削減することの意義を説明したいと思います．例えば，

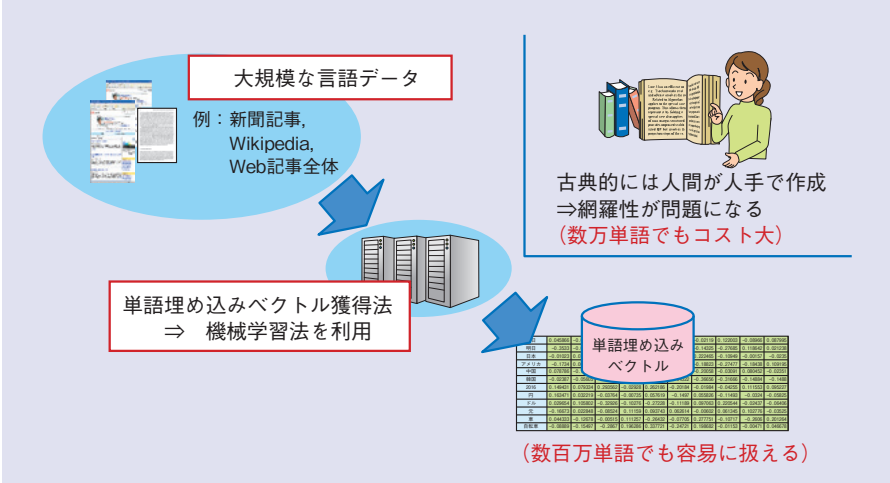


図2　単語埋め込みベクトルの獲得法

人間と対話するロボットのシステムを構築していると仮定します．システムにとって知らない未知の単語が入ってきた場合，それを適切に処理するのは非常に困難となるため，できるだけシステムの知っている語彙数は多くしたいという要求が出ます．そこでシステムに組み込む単語埋め込みベクトルの語彙数をなるべく多くしたくなります．一方で，単語埋め込みベクトルの実体は，各単語に対して1つのD次元のベクトル分の記憶容量を必要とします．つまり，システム（ロボット）にとって「知っている」語彙数を増やせば増やすほど，それを覚えておくために記憶容量を増やさなくてはいけないことになります．例えば，ベクトルの次元数Dが300で，語彙数が300万語彙の場合を考えてみましょう．通常1つの実数値を表現するのに単精度浮動小数なら4バイトの記憶容量が必要となるので，単純計算で，3000000（単語）×300（次元）×4（バイト）=3600000000（バイト）となります．つまり，約3.4 GBの記憶容量が必要，という計算になります．これは，システム全体で3.4 GB必要ということではなく，単語の意味を賢く扱うための1つのモジュールにこれほどの記憶容量が必要になるということです．これは必要記憶容量が大きすぎると思います．

ここで，例えば，必要記憶容量が100分の1の34 MBだったと仮定してみましょう．そうすると，各ロボットに組み込むべき記憶容量を大幅に削減できる可能性が高くなります．今後，一般に対話ロボットが普及するような状況を想定するなら，これらのコスト削減は大きな意味を持つことになります．また別の観点として，3.4 GBでは難しいですが，34 MBなら，スマホなどの携帯端末上のアプリでも十分利用可能といえると思います．仮に今後携帯端末に搭載される記憶容量が激増したとしても，占有使用記憶容量が減れば，それだけ消費電力も軽減されるため，必要記憶容量の削減は省電力にも貢献することができます．このように，単に必要記憶容量を削減するだけの技術でも，さまざまな効果を得られることが期待できます．

■削減方法

具体的な削減方法は，グループ正則化，双対分解，拡張ラグランジュ緩和，クラスタリングといったいくつかの機械学習の技法を駆使して構築されています．やや複雑ですが，ここでは概要を簡潔に説明したいと思います．まず前述のとおり，単語埋め込みベクトルは大量の文章データから，機械学習法を用いて自動的に構築されています．より具体的には，各単語に割り当てられたベクトルの各要素の値を決定する処理が「学習」に相当します．提案法の基本のアイデアとして，各単語のベクトルに同じ値の並びとなる部分がたくさん存在すると，その部分は冗長な情報となるので，1つだけ残してほかは削除し，その代わりに同じ数値の並びがあった情報を付与すれば，情報の欠損は発生しません．このとき，もともとの数値の並びよりも，同じ情報があったという付加情報を記述するのに必要な記憶容量が小さければ，その分，必要記憶容量は削減できるということになります（**図3**）．このアイデアに基づくと数値の並びの総パターン数が少なければ少ないほど単語埋め込みベクトルに必要な記憶容量を減らすことができます．ただし，残念なことに良

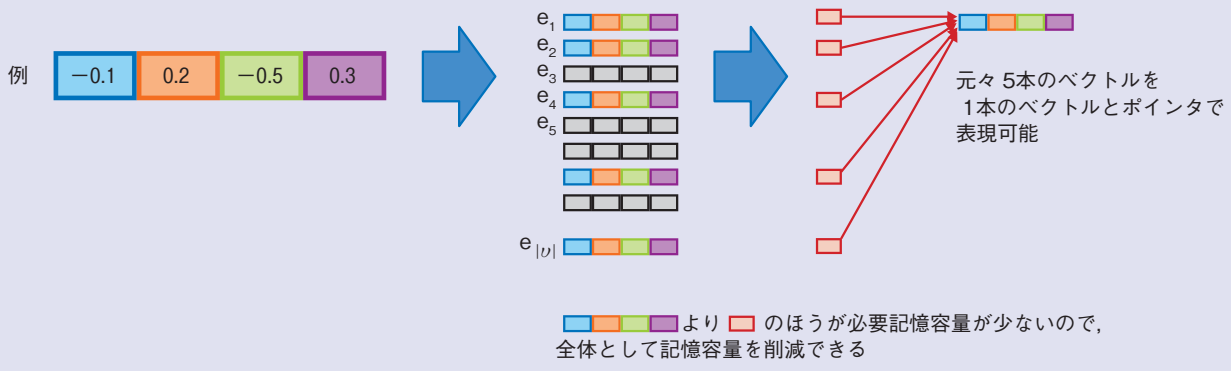
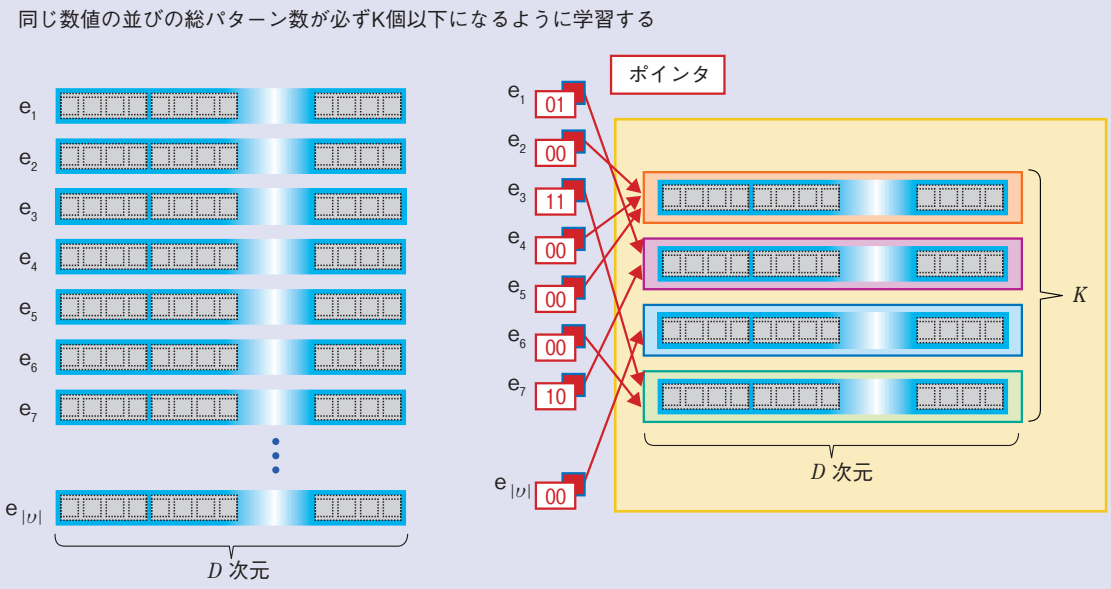


図3　記憶容量の削減法の基本アイデア



(a) 通常の単語埋め込みベクトル　(b) 提案法により記憶容量を減らした埋め込みベクトル

図 4　提案法による単語埋め込みベクトルの必要容量削減

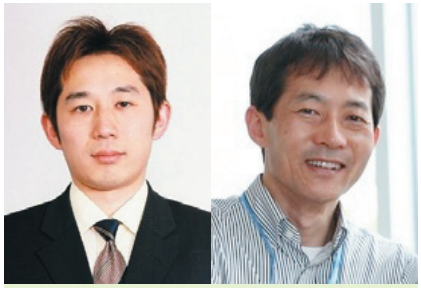
く知られた従来法を用いても，同じ数値の並びのパターンが都合良く得られることは一般的にはありません．そこで，単語埋め込みベクトルを学習する方法論を改良し，「同じ数値の並びの総パターン数が事前に設定したK個以下」になるという条件の下，従来と同じように単語埋め込みベクトルの性能がもっとも高くなるように学習をします．これにより，利用者が所望する記憶容量となる単語埋め込みベクトルを獲得することができるようになります（**図 4**）．

今後の展開

コンピュータやシステムの開発者にとって「使いやすい」技術をつくり出すことにより，最先端の研究成果が単に研究で終わることなく，自然言語を介するAI関連システムで実際に使われる基盤技術となるように研究を進めています．最終的には，あらゆるAI関連システムで使われるような汎用性と利便性が高い技術をめざしています．また，本技術を継続的に発展させることで，利用している実システムの性能や性質の底上げも可能であることが考えられます．

■参考文献

(1) G.E. Hinton：“Learning Distributed Representations of Concepts,” Proc. of the Eighth Annual Conference of the Cognitive Science Society, pp.1-12, Amherst, U.S.A., August 1986.
(2) T. Mikolov, K. Chen, G. Corrado, and J. Dean：“Efficient Estimation of Word Representations in Vector Space,” Proc. of ICLR 2013, Scottsdale, U.S.A., May 2013.
(3) J. Suzuki and M. Nagata：“A Unified Learning Framework of Skip-Grams and Global Vectors,” Proc. of ACL-ICNLP2015, Beijing, China, July 2015.
(4) J. Suzuki and M. Nagata：“Right-truncatable Neural Word Embeddings,” Proc. of NAACL-HLT2016, San Diego, U.S.A., June 2016.
(5) J. Suzuki and M. Nagata：“Learning Compact Neural Word Embeddings by Parameter Space Sharing,” Proc. of IJCAI-16, New York, U.S.A., July 2016.



（左から）鈴木　潤/ 永田 昌明

コンピュータによる自然言語の理解は，AI技術の中でも最難関の課題の1つであり，かつ，実現した際の社会的なインパクトは計り知れない最重要技術と考えられています．今後も着実に研究を重ね，早期の実現をめざしていきます．

◆問い合わせ先
NTTコミュニケーション科学基礎研究所
協創情報研究部
TEL　0774-93-5361
FAX　0774-93-5385
E-mail　suzuki.jun@lab.ntt.co.jp